Weekly Report

Lu Junhua

2015 年 7 月 12 日

On Monday, we had a discussion with TGRAM company, and identify several problems and tasks. We had a rough knowledge of all the data. Later, Ke, Feng and I had several discussion to configure all the data attribute we need. They are shown below(on page2).

After that, Prof. He made several suggestions on sampling of data.

- Compute crime rate firstly. And we would sample the data according to the rate.(However, we had claim the problem of too low crime rate, maybe we should change the way of crime rate.)
- Guarantee of random sampling. Both in criminals and normal persons.
- Sampling by household(按户抽样). That means, once someone is sampled, all the information of all the people under same households should be extracted.

As He said, with the several suggestions, we can figure out the best-approximated results to the result came out from processing all the data used. I also consulted with Prof. Gu, since previously all the people we used last time are all with hotel records. And this time, we will not be limited by this constraint. If one without internet cafe/hotel records, we fill it as zero.

And next week, we will move our computer in TGRAM and our group to Zhongshan Building. All the data will be processed there, since Gongan do not want any classified data revealing.

Besides this, I learned more about Numpy in python, a tool for scientific computing, data processing. I came across several problems while understanding the functions, but I managed to figure it out. I also learned more about SVM on coursera. While the description of models on Tongji Xuexi Fangfa is too sketchy, the video on coursera is much more complete. Many details are overlooked on the book, like $\min = 1$ and all ≥ 1 are not equal constraint in basic svm model, but the can reach the same result and this conclusion needed proving, but this details is not mentioned in the blue book. I think if we are to master one method, we must spent more time on these details.

SVM and its applications can be classified into linear support vector machine, dual support vector machine, kernel support vector machine, soft-margin support vector machine, kernel logistic regression and support vector regression. I will learn it step by step.

Next week, I will mostly contribute myself on processing the data in Zhongshan Building. I will go home as planned. Also, Prof He says he will be Hangzhou on the beginning and ending of his stay in China(25th July to 12th August), so I planned to have a meet with him on August as I return Hangzhou.

- 2 万个犯罪人员, 10 万个未犯罪人员.
  - 个人基本数据
    - 年龄
    - 身份证号(必须)
    - 性别
    - 出生地(籍贯 6 位)
    - 婚姻
    - 文化程度
    - 职业
    - 是否犯罪前科，犯罪次数和细节(包括时间，地点，犯罪类别等等)
    - 户号
    - 父母儿女犯罪次数(看看能不能看到兄弟姐妹) (这里又是关于户号这一个问题, 下面有讲)
  - 正常人一定是要在现有记录中都没有任何犯罪记录的人.
  - 犯罪人员中要有一定比例的累案犯.
  - 一定要有身份证这个信息以确保以后想使用新的字段时候能够对号入座
  - 之前职业，婚姻等数据中包含很多缺失数据，希望能够尽量减少
  - 常住人口、暂住人口比例问题，暂定 1:1

- 犯罪记录的详细程度
  - 对于犯罪人员，最好将其犯罪信息都给我们,诸如:
    - 犯罪时间
    - 犯罪内容(类别,如偷窃,吸毒)
    - 犯罪次数(有记录即可,我们可以自己统计)

- 网吧记录,开房记录
  - 所有人的详细的网吧上网记录，旅馆住宿记录. 至少要有每个月详细数据,明确开始和结束的时间,没有的话至少要有次数
  - 我们对网吧,开房地点这个变量有兴趣. (先弄下来把)

- 籍贯的地理位置数据
  - 施总说有一张精确到县区的常用高发作案方法手段统计图,这个很重要
  - 全国各地经纬度表网上是有的，然而我发现不同版本还有轻微出入，并且未发现官方的文档. 如果有，可否提供一张身份证前六位和经纬度直接对应的全面官方的表格?以我经验，我家乡在由县变市之前，身份证前六位是 320626(我父母辈及以上都是这样的)，而变县级市之后变成了 320681，而网上大多不保留县，甚至有些地方会改名(云南思茅 07 年改名普洱)等，因而我不保证这些地方数据的准确性或者说能不能对上号.

- 农业/非农业户口

- 有/无机动车